autogluon_each_location

October 9, 2023

[6]:
```
# config

label = 'y'
metric = 'mean_absolute_error'
time_limit = 60*30
presets = 'best_quality'

do_drop_ds = True

use_groups = False
n_groups = 8

auto_stack = True
num_stack_levels = 1
num_bag_folds = 0
if auto_stack:
    num_stack_levels = None
    num_bag_folds = None

use_tune_data = False
use_test_data = True
tune_and_test_length = 24*30*3 # 3 months from end, this changes the
↪evaluations for only test
holdout_frac = None
use_bag_holdout = False # Enable this if there is a large gap between score_val
↪and score_test in stack models.

sample_weight = 'sample_weight' #None
weight_evaluation = True #False
sample_weight_estimated = 3 # this changes evaluations for test and tune WTF,
↪cant find a fix

run_analysis = False
```

[7]:
```
import pandas as pd
import numpy as np
```

```python
import warnings
warnings.filterwarnings("ignore")

def fix_datetime(X, name):
    # Convert 'date_forecast' to datetime format and replace original column␣
↪with 'ds'
    X['ds'] = pd.to_datetime(X['date_forecast'])
    X.drop(columns=['date_forecast'], inplace=True, errors='ignore')
    X.sort_values(by='ds', inplace=True)
    X.set_index('ds', inplace=True)

    # Drop rows where the minute part of the time is not 0
    X = X[X.index.minute == 0].copy()
    return X



def convert_to_datetime(X_train_observed, X_train_estimated, X_test, y_train):
    X_train_observed = fix_datetime(X_train_observed, "X_train_observed")
    X_train_estimated = fix_datetime(X_train_estimated, "X_train_estimated")
    X_test = fix_datetime(X_test, "X_test")

    # add sample weights, which are 1 for observed and 3 for estimated
    X_train_observed["sample_weight"] = 1
    X_train_estimated["sample_weight"] = sample_weight_estimated
    X_test["sample_weight"] = sample_weight_estimated

    X_train_observed["estimated_diff_hours"] = 0
    X_train_estimated["estimated_diff_hours"] = (X_train_estimated.index - pd.
↪to_datetime(X_train_estimated["date_calc"])).dt.total_seconds() / 3600
    X_test["estimated_diff_hours"] = (X_test.index - pd.
↪to_datetime(X_test["date_calc"])).dt.total_seconds() / 3600

    X_train_estimated["estimated_diff_hours"] =␣
↪X_train_estimated["estimated_diff_hours"].astype('int64')
    # the filled once will get dropped later anyways, when we drop y nans
    X_test["estimated_diff_hours"] = X_test["estimated_diff_hours"].fillna(-50).
↪astype('int64')


    X_train_estimated.drop(columns=['date_calc'], inplace=True)
    X_test.drop(columns=['date_calc'], inplace=True)

    y_train['ds'] = pd.to_datetime(y_train['time'])
    y_train.drop(columns=['time'], inplace=True)
```

```
    y_train.sort_values(by='ds', inplace=True)
    y_train.set_index('ds', inplace=True)

    return X_train_observed, X_train_estimated, X_test, y_train




def preprocess_data(X_train_observed, X_train_estimated, X_test, y_train, location):
    # convert to datetime
    X_train_observed, X_train_estimated, X_test, y_train = convert_to_datetime(X_train_observed, X_train_estimated, X_test, y_train)

    y_train["y"] = y_train["pv_measurement"].astype('float64')
    y_train.drop(columns=['pv_measurement'], inplace=True)
    X_train = pd.concat([X_train_observed, X_train_estimated])

    # fill missng sample_weight with 3
    #X_train["sample_weight"] = X_train["sample_weight"].fillna(0)


    # clip all y values to 0 if negative
    y_train["y"] = y_train["y"].clip(lower=0)

    X_train = pd.merge(X_train, y_train, how="inner", left_index=True, right_index=True)

    # print number of nans in sample_weight
    print(f"Number of nans in sample_weight: {X_train['sample_weight'].isna().sum()}")
    # print number of nans in y
    print(f"Number of nans in y: {X_train['y'].isna().sum()}")


    X_train["location"] = location
    X_test["location"] = location

    return X_train, X_test
# Define locations
locations = ['A', 'B', 'C']

X_trains = []
X_tests = []
# Loop through locations
for loc in locations:
    print(f"Processing location {loc}...")
```

```
    # Read target training data
    y_train = pd.read_parquet(f'{loc}/train_targets.parquet')

    # Read estimated training data and add location feature
    X_train_estimated = pd.read_parquet(f'{loc}/X_train_estimated.parquet')

    # Read observed training data and add location feature
    X_train_observed= pd.read_parquet(f'{loc}/X_train_observed.parquet')

    # Read estimated test data and add location feature
    X_test_estimated = pd.read_parquet(f'{loc}/X_test_estimated.parquet')

    # Preprocess data
    X_train, X_test = preprocess_data(X_train_observed, X_train_estimated,␣
 ↪X_test_estimated, y_train, loc)

    X_trains.append(X_train)
    X_tests.append(X_test)

# Concatenate all data and save to csv
X_train = pd.concat(X_trains)
X_test = pd.concat(X_tests)
```

```
Processing location A…
Number of nans in sample_weight: 0
Number of nans in y: 0
Processing location B…
Number of nans in sample_weight: 0
Number of nans in y: 4
Processing location C…
Number of nans in sample_weight: 0
Number of nans in y: 6059
```

1 Feature enginering

[8]:
```
import numpy as np
import pandas as pd

X_train.dropna(subset=['y'], inplace=True)

if not do_drop_ds:
    # add hour datetime feature
    X_train["hour"] = X_train.index.hour
    X_test["hour"] = X_test.index.hour

#print(X_train.head())
```

```
if use_groups:
    # fix groups for cross validation
    locations = X_train['location'].unique()  # Assuming 'location' is the name␣
 ↪of the column representing locations

    grouped_dfs = []  # To store data frames split by location

    # Loop through each unique location
    for loc in locations:
        loc_df = X_train[X_train['location'] == loc]

        # Sort the DataFrame for this location by the time column
        loc_df = loc_df.sort_index()

        # Calculate the size of each group for this location
        group_size = len(loc_df) // n_groups

        # Create a new 'group' column for this location
        loc_df['group'] = np.repeat(range(n_groups),␣
 ↪repeats=[group_size]*(n_groups-1) + [len(loc_df) - group_size*(n_groups-1)])

        # Append to list of grouped DataFrames
        grouped_dfs.append(loc_df)

    # Concatenate all the grouped DataFrames back together
    X_train = pd.concat(grouped_dfs)
    X_train.sort_index(inplace=True)
    print(X_train["group"].head())





to_drop = ["snow_drift:idx", "snow_density:kgm3"]

X_train.drop(columns=to_drop, inplace=True)
X_test.drop(columns=to_drop, inplace=True)

X_train.to_csv('X_train_raw.csv', index=True)
X_test.to_csv('X_test_raw.csv', index=True)
```

[9]:
```
from autogluon.tabular import TabularDataset, TabularPredictor
from autogluon.timeseries import TimeSeriesDataFrame
import numpy as np
train_data = TabularDataset('X_train_raw.csv')
```

```
# set group column of train_data be increasing from 0 to 7 based on time, the␣
↪first 1/8 of the data is group 0, the second 1/8 of the data is group 1, etc.
train_data['ds'] = pd.to_datetime(train_data['ds'])
train_data = train_data.sort_values(by='ds')

# # print size of the group for each location
# for loc in locations:
#     print(f"Location {loc}:")
#     print(train_data[train_data["location"] == loc].groupby('group').size())


# get end date of train data and subtract 3 months
split_time = pd.to_datetime(train_data["ds"]).max() - pd.
↪Timedelta(hours=tune_and_test_length)
train_set = TabularDataset(train_data[train_data["ds"] < split_time])
test_set = TabularDataset(train_data[train_data["ds"] >= split_time])
if use_groups:
    test_set = test_set.drop(columns=['group'])

if do_drop_ds:
    train_set = train_set.drop(columns=['ds'])
    test_set = test_set.drop(columns=['ds'])
    train_data = train_data.drop(columns=['ds'])


def normalize_sample_weights_per_location(df):
    for loc in locations:
        loc_df = df[df["location"] == loc]
        loc_df["sample_weight"] = loc_df["sample_weight"] /␣
↪loc_df["sample_weight"].sum() * loc_df.shape[0]
        df[df["location"] == loc] = loc_df
    return df

tuning_data = None
if use_tune_data:
    train_data = train_set
    if use_test_data:
        # split test_set in half, use first half for tuning
        tuning_data, test_data = [], []
        for loc in locations:
            loc_test_set = test_set[test_set["location"] == loc]
            loc_tuning_data = loc_test_set.iloc[:len(loc_test_set)//2]
            loc_test_data = loc_test_set.iloc[len(loc_test_set)//2:]
            tuning_data.append(loc_tuning_data)
            test_data.append(loc_test_data)
        tuning_data = pd.concat(tuning_data)
        test_data = pd.concat(test_data)
```

```
        print("Shapes of tuning and test", tuning_data.shape[0], test_data.
 ↪shape[0], tuning_data.shape[0] + test_data.shape[0])

    else:
        tuning_data = test_set
        print("Shape of tuning", tuning_data.shape[0])

    # ensure sample weights for your tuning data sum to the number of rows in␣
 ↪the tuning data.
    tuning_data = normalize_sample_weights_per_location(tuning_data)


else:
    if use_test_data:
        train_data = train_set
        test_data = test_set
        print("Shape of test", test_data.shape[0])

# ensure sample weights for your training (or tuning) data sum to the number of␣
 ↪rows in the training (or tuning) data.
train_data = normalize_sample_weights_per_location(train_data)
if use_test_data:
    test_data = normalize_sample_weights_per_location(test_data)
```

```
Shape of test 5791
```

[10]:
```
if run_analysis:
    import autogluon.eda.auto as auto
    auto.dataset_overview(train_data=train_data, test_data=test_data,␣
 ↪label="y", sample=None)
```

[11]:
```
if run_analysis:
    auto.target_analysis(train_data=train_data, label="y")
```

2 Starting

[12]:
```
import os


# Get the last submission number
last_submission_number = int(max([int(filename.split('_')[1].split('.')[0]) for␣
 ↪filename in os.listdir('submissions') if "submission" in filename]))
print("Last submission number:", last_submission_number)
print("Now creating submission number:", last_submission_number + 1)

# Create the new filename
new_filename = f'submission_{last_submission_number + 1}'
```

```
hello = os.environ.get('HELLO')
if hello is not None:
    new_filename += f'_{hello}'

print("New filename:", new_filename)
```

```
Last submission number: 82
Now creating submission number: 83
New filename: submission_83
```

[13]:
```
predictors = [None, None, None]
```

[14]:
```
def fit_predictor_for_location(loc):
    print(f"Training model for location {loc}...")
    # sum of sample weights for this location, and number of rows, for both␣
 ↪train and tune data and test data
    print("Train data sample weight sum:", train_data[train_data["location"] ==␣
 ↪loc]["sample_weight"].sum())
    print("Train data number of rows:", train_data[train_data["location"] ==␣
 ↪loc].shape[0])
    if use_tune_data:
        print("Tune data sample weight sum:",␣
 ↪tuning_data[tuning_data["location"] == loc]["sample_weight"].sum())
        print("Tune data number of rows:", tuning_data[tuning_data["location"]␣
 ↪== loc].shape[0])
    if use_test_data:
        print("Test data sample weight sum:", test_data[test_data["location"]␣
 ↪== loc]["sample_weight"].sum())
        print("Test data number of rows:", test_data[test_data["location"] ==␣
 ↪loc].shape[0])
    predictor = TabularPredictor(
        label=label,
        eval_metric=metric,
        path=f"AutogluonModels/{new_filename}_{loc}",
        sample_weight=sample_weight,
        weight_evaluation=weight_evaluation,
        groups="group" if use_groups else None,
    ).fit(
        train_data=train_data[train_data["location"] == loc],
        time_limit=time_limit,
        #presets=presets,
        num_stack_levels=num_stack_levels,
        num_bag_folds=num_bag_folds if not use_groups else 2,# just put␣
 ↪somethin, will be overwritten anyways
        tuning_data=tuning_data[tuning_data["location"] == loc] if␣
 ↪use_tune_data else None,
```

```
        use_bag_holdout=use_bag_holdout,
        holdout_frac=holdout_frac,
    )

    # evaluate on test data
    if use_test_data:
        # drop sample_weight column
        t = test_data[test_data["location"] == loc]#.
↪drop(columns=["sample_weight"])
        perf = predictor.evaluate(t)
        print("Evaluation on test data:")
        print(perf[predictor.eval_metric.name])

    return predictor

loc = "A"
predictors[0] = fit_predictor_for_location(loc)
```

```
Training model for location A…
Train data sample weight sum: 31899.999999999996
```

```
Values in column 'sample_weight' used as sample weights instead of predictive
features. Evaluation will report weighted metrics, so ensure same column exists
in test data.
Beginning AutoGluon training … Time limit = 1800s
AutoGluon will save models to "AutogluonModels/submission_83_A/"
AutoGluon Version:  0.8.2
Python Version:     3.10.12
Operating System:   Linux
Platform Machine:   x86_64
Platform Version:   #1 SMP Debian 5.10.197-1 (2023-09-29)
Disk Space Avail:   304.26 GB / 315.93 GB (96.3%)
Train Data Rows:    31900
Train Data Columns: 46
Label Column: y
Preprocessing data …
AutoGluon infers your prediction problem is: 'regression' (because dtype of
label-column == float and many unique label-values observed).
        Label info (max, min, mean, stddev): (5733.42, 0.0, 633.132, 1165.64686)
        If 'regression' is not the correct problem_type, please manually specify
the problem_type parameter during predictor init (You may specify problem_type
as one of: ['binary', 'multiclass', 'regression'])
Using Feature Generators to preprocess the data …
Fitting AutoMLPipelineFeatureGenerator…
        Available Memory:                    132461.02 MB
        Train Data (Original)  Memory Usage: 13.08 MB (0.0% of available memory)
        Inferring data type of each feature based on column values. Set
feature_metadata_in to manually specify special dtypes of the features.
```

```
        Stage 1 Generators:
                Fitting AsTypeFeatureGenerator…
                        Note: Converting 3 features to boolean dtype as they
only contain 2 unique values.
        Stage 2 Generators:
                Fitting FillNaFeatureGenerator…
        Stage 3 Generators:
                Fitting IdentityFeatureGenerator…
        Stage 4 Generators:
                Fitting DropUniqueFeatureGenerator…
```

```
Train data number of rows: 31900
Test data sample weight sum: 2161
Test data number of rows: 2161
```

```
        Stage 5 Generators:
                Fitting DropDuplicatesFeatureGenerator…
        Useless Original Features (Count: 2): ['elevation:m', 'location']
                These features carry no predictive signal and should be manually
investigated.
                This is typically a feature which has the same value for all
rows.
                These features do not need to be present at inference time.
        Types of features in original data (raw dtype, special dtypes):
                ('float', []) : 42 | ['absolute_humidity_2m:gm3',
'air_density_2m:kgm3', 'ceiling_height_agl:m', 'clear_sky_energy_1h:J',
'clear_sky_rad:W', …]
                ('int', [])   :  1 | ['estimated_diff_hours']
        Types of features in processed data (raw dtype, special dtypes):
                ('float', [])     : 39 | ['absolute_humidity_2m:gm3',
'air_density_2m:kgm3', 'ceiling_height_agl:m', 'clear_sky_energy_1h:J',
'clear_sky_rad:W', …]
                ('int', [])       :  1 | ['estimated_diff_hours']
                ('int', ['bool']) :  3 | ['is_day:idx', 'is_in_shadow:idx',
'wind_speed_w_1000hPa:ms']
        0.2s = Fit runtime
        43 features in original data used to generate 43 features in processed
data.
        Train Data (Processed) Memory Usage: 10.3 MB (0.0% of available memory)
Data preprocessing and feature engineering runtime = 0.2s …
AutoGluon will gauge predictive performance using evaluation metric:
'mean_absolute_error'
        This metric's sign has been flipped to adhere to being higher_is_better.
The metric score can be multiplied by -1 to get the metric value.
        To change this, specify the eval_metric parameter of Predictor()
Automatically generating train/validation split with
holdout_frac=0.07836990595611286, Train Rows: 29400, Val Rows: 2500
User-specified model hyperparameters to be fit:
{
```

```
        'NN_TORCH': {},
        'GBM': [{'extra_trees': True, 'ag_args': {'name_suffix': 'XT'}}, {},
'GBMLarge'],
        'CAT': {},
        'XGB': {},
        'FASTAI': {},
        'RF': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'XT': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'KNN': [{'weights': 'uniform', 'ag_args': {'name_suffix': 'Unif'}},
{'weights': 'distance', 'ag_args': {'name_suffix': 'Dist'}}],
}
Fitting 11 L1 models …
Fitting model: KNeighborsUnif … Training model for up to 1799.8s of the
1799.79s of remaining time.
        -256.4362        = Validation score   (-mean_absolute_error)
        0.04s    = Training   runtime
        0.08s    = Validation runtime
Fitting model: KNeighborsDist … Training model for up to 1799.67s of the
1799.66s of remaining time.
        -258.904         = Validation score   (-mean_absolute_error)
        0.04s    = Training   runtime
        0.04s    = Validation runtime
Fitting model: LightGBMXT … Training model for up to 1799.58s of the 1799.58s
of remaining time.

[1000]  valid_set's l1: 160.225
[2000]  valid_set's l1: 156.638
[3000]  valid_set's l1: 154.856
[4000]  valid_set's l1: 153.333
[5000]  valid_set's l1: 152.53
[6000]  valid_set's l1: 152.107
[7000]  valid_set's l1: 151.576
[8000]  valid_set's l1: 150.955
[9000]  valid_set's l1: 150.397
[10000] valid_set's l1: 150.014

        -149.9836        = Validation score   (-mean_absolute_error)
        14.7s    = Training   runtime
        0.19s    = Validation runtime
Fitting model: LightGBM … Training model for up to 1784.31s of the 1784.31s of
remaining time.
```

```
[1000]  valid_set's l1: 164.748
[2000]  valid_set's l1: 163.139
[3000]  valid_set's l1: 161.84
[4000]  valid_set's l1: 161.376
[5000]  valid_set's l1: 161.17
[6000]  valid_set's l1: 161.09
[7000]  valid_set's l1: 161.039
[8000]  valid_set's l1: 161.06

        -160.9804        = Validation score   (-mean_absolute_error)
        13.12s   = Training   runtime
        0.12s    = Validation runtime
Fitting model: RandomForestMSE … Training model for up to 1770.89s of the
1770.88s of remaining time.
        -168.3222        = Validation score   (-mean_absolute_error)
        8.31s    = Training   runtime
        0.09s    = Validation runtime
Fitting model: CatBoost … Training model for up to 1762.02s of the 1762.01s of
remaining time.
        -164.7748        = Validation score   (-mean_absolute_error)
        114.88s  = Training   runtime
        0.01s    = Validation runtime
Fitting model: ExtraTreesMSE … Training model for up to 1647.1s of the
1647.09s of remaining time.
        -168.0615        = Validation score   (-mean_absolute_error)
        1.76s    = Training   runtime
        0.09s    = Validation runtime
Fitting model: NeuralNetFastAI … Training model for up to 1644.77s of the
1644.76s of remaining time.
        -173.685         = Validation score   (-mean_absolute_error)
        27.5s    = Training   runtime
        0.04s    = Validation runtime
Fitting model: XGBoost … Training model for up to 1617.19s of the 1617.18s of
remaining time.
        -166.9422        = Validation score   (-mean_absolute_error)
        2.95s    = Training   runtime
        0.02s    = Validation runtime
Fitting model: NeuralNetTorch … Training model for up to 1614.2s of the
1614.19s of remaining time.
        -157.4766        = Validation score   (-mean_absolute_error)
        53.37s   = Training   runtime
        0.04s    = Validation runtime
Fitting model: LightGBMLarge … Training model for up to 1560.78s of the
1560.77s of remaining time.

[1000]  valid_set's l1: 152.501
[2000]  valid_set's l1: 151.16
[3000]  valid_set's l1: 150.868
[4000]  valid_set's l1: 150.796
```

```
[5000]  valid_set's l1: 150.752
[6000]  valid_set's l1: 150.73
[7000]  valid_set's l1: 150.721
[8000]  valid_set's l1: 150.718
[9000]  valid_set's l1: 150.715
[10000] valid_set's l1: 150.715
        -150.7145        = Validation score   (-mean_absolute_error)
        44.88s   = Training   runtime
        0.27s    = Validation runtime
Fitting model: WeightedEnsemble_L2 … Training model for up to 360.0s of the
1514.44s of remaining time.
        -144.7512        = Validation score   (-mean_absolute_error)
        0.47s    = Training   runtime
        0.0s     = Validation runtime
AutoGluon training complete, total runtime = 286.08s … Best model:
"WeightedEnsemble_L2"
TabularPredictor saved. To load, use: predictor =
TabularPredictor.load("AutogluonModels/submission_83_A/")
WARNING: eval_metric='pearsonr' does not support sample weights so they will be
ignored in reported metric.
Evaluation: mean_absolute_error on test data: -188.93721938998613
        Note: Scores are always higher_is_better. This metric score can be
multiplied by -1 to get the metric value.
Evaluations on test data:
{
    "mean_absolute_error": -188.93721938998613,
    "root_mean_squared_error": -412.50206821119434,
    "mean_squared_error": -170157.9562785128,
    "r2": 0.876539550093958,
    "pearsonr": 0.9364619268059786,
    "median_absolute_error": -11.871703147888184
}

Evaluation on test data:
-188.93721938998613
```

[15]:
```
loc = "B"
predictors[1] = fit_predictor_for_location(loc)
```

```
Values in column 'sample_weight' used as sample weights instead of predictive
features. Evaluation will report weighted metrics, so ensure same column exists
in test data.
Beginning AutoGluon training … Time limit = 1800s
AutoGluon will save models to "AutogluonModels/submission_83_B/"
AutoGluon Version:  0.8.2
Python Version:     3.10.12
Operating System:   Linux
Platform Machine:   x86_64
```

```
Platform Version:   #1 SMP Debian 5.10.197-1 (2023-09-29)
Disk Space Avail:   303.27 GB / 315.93 GB (96.0%)
Train Data Rows:    30768
Train Data Columns: 46
Label Column: y
Preprocessing data …
AutoGluon infers your prediction problem is: 'regression' (because dtype of
label-column == float and many unique label-values observed).
        Label info (max, min, mean, stddev): (1152.3, -0.0, 97.74541, 195.0957)
        If 'regression' is not the correct problem_type, please manually specify
the problem_type parameter during predictor init (You may specify problem_type
as one of: ['binary', 'multiclass', 'regression'])
Using Feature Generators to preprocess the data …
Fitting AutoMLPipelineFeatureGenerator…
        Available Memory:                    130687.49 MB
        Train Data (Original)  Memory Usage: 12.62 MB (0.0% of available memory)
        Inferring data type of each feature based on column values. Set
feature_metadata_in to manually specify special dtypes of the features.
        Stage 1 Generators:
                Fitting AsTypeFeatureGenerator…
                        Note: Converting 3 features to boolean dtype as they
only contain 2 unique values.
        Stage 2 Generators:
                Fitting FillNaFeatureGenerator…
        Stage 3 Generators:
                Fitting IdentityFeatureGenerator…
        Stage 4 Generators:
                Fitting DropUniqueFeatureGenerator…

Training model for location B…
Train data sample weight sum: 30768.0
Train data number of rows: 30768
Test data sample weight sum: 2051
Test data number of rows: 2051

        Stage 5 Generators:
                Fitting DropDuplicatesFeatureGenerator…
        Useless Original Features (Count: 2): ['elevation:m', 'location']
                These features carry no predictive signal and should be manually
investigated.
                This is typically a feature which has the same value for all
rows.
                These features do not need to be present at inference time.
        Types of features in original data (raw dtype, special dtypes):
                ('float', []) : 42 | ['absolute_humidity_2m:gm3',
'air_density_2m:kgm3', 'ceiling_height_agl:m', 'clear_sky_energy_1h:J',
'clear_sky_rad:W', …]
                ('int', [])   :  1 | ['estimated_diff_hours']
        Types of features in processed data (raw dtype, special dtypes):
```

```
                ('float', [])     : 39 | ['absolute_humidity_2m:gm3',
'air_density_2m:kgm3', 'ceiling_height_agl:m', 'clear_sky_energy_1h:J',
'clear_sky_rad:W', …]
                ('int', [])       :  1 | ['estimated_diff_hours']
                ('int', ['bool']) :  3 | ['is_day:idx', 'is_in_shadow:idx',
'wind_speed_w_1000hPa:ms']
        0.2s = Fit runtime
        43 features in original data used to generate 43 features in processed
data.
        Train Data (Processed) Memory Usage: 9.94 MB (0.0% of available memory)
Data preprocessing and feature engineering runtime = 0.2s …
AutoGluon will gauge predictive performance using evaluation metric:
'mean_absolute_error'
        This metric's sign has been flipped to adhere to being higher_is_better.
The metric score can be multiplied by -1 to get the metric value.
        To change this, specify the eval_metric parameter of Predictor()
Automatically generating train/validation split with
holdout_frac=0.0812532501300052, Train Rows: 28268, Val Rows: 2500
User-specified model hyperparameters to be fit:
{
        'NN_TORCH': {},
        'GBM': [{'extra_trees': True, 'ag_args': {'name_suffix': 'XT'}}, {},
'GBMLarge'],
        'CAT': {},
        'XGB': {},
        'FASTAI': {},
        'RF': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'XT': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'KNN': [{'weights': 'uniform', 'ag_args': {'name_suffix': 'Unif'}},
{'weights': 'distance', 'ag_args': {'name_suffix': 'Dist'}}],
}
Fitting 11 L1 models …
Fitting model: KNeighborsUnif … Training model for up to 1799.8s of the
1799.8s of remaining time.
        -53.446  = Validation score   (-mean_absolute_error)
        0.04s    = Training   runtime
        0.04s    = Validation runtime
Fitting model: KNeighborsDist … Training model for up to 1799.72s of the
1799.71s of remaining time.
        -53.26   = Validation score   (-mean_absolute_error)
```

```
        0.04s    = Training   runtime
        0.04s    = Validation runtime
Fitting model: LightGBMXT … Training model for up to 1799.63s of the 1799.63s
of remaining time.
[1000]  valid_set's l1: 33.1184
[2000]  valid_set's l1: 31.2902
[3000]  valid_set's l1: 30.3861
[4000]  valid_set's l1: 29.7745
[5000]  valid_set's l1: 29.2881
[6000]  valid_set's l1: 28.9165
[7000]  valid_set's l1: 28.69
[8000]  valid_set's l1: 28.5065
[9000]  valid_set's l1: 28.3414
[10000] valid_set's l1: 28.1891
        -28.1891         = Validation score   (-mean_absolute_error)
        13.87s   = Training   runtime
        0.18s    = Validation runtime
Fitting model: LightGBM … Training model for up to 1785.29s of the 1785.29s of
remaining time.
[1000]  valid_set's l1: 30.9525
[2000]  valid_set's l1: 29.5944
[3000]  valid_set's l1: 28.9858
[4000]  valid_set's l1: 28.7098
[5000]  valid_set's l1: 28.5117
[6000]  valid_set's l1: 28.3986
[7000]  valid_set's l1: 28.3279
[8000]  valid_set's l1: 28.2614
[9000]  valid_set's l1: 28.2242
[10000] valid_set's l1: 28.2037
        -28.2036         = Validation score   (-mean_absolute_error)
        14.2s    = Training   runtime
        0.17s    = Validation runtime
Fitting model: RandomForestMSE … Training model for up to 1770.63s of the
1770.62s of remaining time.
        -32.8223         = Validation score   (-mean_absolute_error)
        9.33s    = Training   runtime
        0.1s     = Validation runtime
Fitting model: CatBoost … Training model for up to 1760.83s of the 1760.82s of
remaining time.
        -30.4129         = Validation score   (-mean_absolute_error)
        115.23s  = Training   runtime
        0.01s    = Validation runtime
Fitting model: ExtraTreesMSE … Training model for up to 1645.56s of the
1645.55s of remaining time.
        -33.619  = Validation score   (-mean_absolute_error)
        1.89s    = Training   runtime
```

```
        0.09s    = Validation runtime
Fitting model: NeuralNetFastAI … Training model for up to 1643.17s of the
1643.16s of remaining time.
        -37.5693         = Validation score   (-mean_absolute_error)
        26.48s   = Training   runtime
        0.04s    = Validation runtime
Fitting model: XGBoost … Training model for up to 1616.62s of the 1616.61s of
remaining time.
        -31.0815         = Validation score   (-mean_absolute_error)
        24.5s    = Training   runtime
        0.21s    = Validation runtime
Fitting model: NeuralNetTorch … Training model for up to 1591.76s of the
1591.75s of remaining time.
        -31.5314         = Validation score   (-mean_absolute_error)
        103.34s  = Training   runtime
        0.04s    = Validation runtime
Fitting model: LightGBMLarge … Training model for up to 1488.37s of the
1488.36s of remaining time.

[1000]  valid_set's l1: 28.0771
[2000]  valid_set's l1: 27.2322
[3000]  valid_set's l1: 27.0078
[4000]  valid_set's l1: 26.931
[5000]  valid_set's l1: 26.9056
[6000]  valid_set's l1: 26.8914
[7000]  valid_set's l1: 26.8859
[8000]  valid_set's l1: 26.8825
[9000]  valid_set's l1: 26.8809
[10000] valid_set's l1: 26.88

        -26.88   = Validation score   (-mean_absolute_error)
        51.57s   = Training   runtime
        0.39s    = Validation runtime
Fitting model: WeightedEnsemble_L2 … Training model for up to 360.0s of the
1435.05s of remaining time.
        -26.5096         = Validation score   (-mean_absolute_error)
        0.45s    = Training   runtime
        0.0s     = Validation runtime
AutoGluon training complete, total runtime = 365.44s … Best model:
"WeightedEnsemble_L2"
TabularPredictor saved. To load, use: predictor =
TabularPredictor.load("AutogluonModels/submission_83_B/")
WARNING: eval_metric='pearsonr' does not support sample weights so they will be
ignored in reported metric.
Evaluation: mean_absolute_error on test data: -37.193940933616126
        Note: Scores are always higher_is_better. This metric score can be
multiplied by -1 to get the metric value.
Evaluations on test data:
{
```

```
    "mean_absolute_error": -37.193940933616126,
    "root_mean_squared_error": -81.67094023712487,
    "mean_squared_error": -6670.1424792160215,
    "r2": 0.7854684374345835,
    "pearsonr": 0.9089080478259688,
    "median_absolute_error": -8.027046203613281
}

Evaluation on test data:
-37.193940933616126
```

[16]:
```python
loc = "C"
predictors[2] = fit_predictor_for_location(loc)
```

```
Values in column 'sample_weight' used as sample weights instead of predictive
features. Evaluation will report weighted metrics, so ensure same column exists
in test data.
Beginning AutoGluon training … Time limit = 1800s
AutoGluon will save models to "AutogluonModels/submission_83_C/"
AutoGluon Version:  0.8.2
Python Version:     3.10.12
Operating System:   Linux
Platform Machine:   x86_64
Platform Version:   #1 SMP Debian 5.10.197-1 (2023-09-29)
Disk Space Avail:   302.33 GB / 315.93 GB (95.7%)
Train Data Rows:    24492
Train Data Columns: 46
Label Column: y
Preprocessing data …
AutoGluon infers your prediction problem is: 'regression' (because dtype of
label-column == float and label-values can't be converted to int).
        Label info (max, min, mean, stddev): (999.6, 0.0, 78.11911, 167.50151)
        If 'regression' is not the correct problem_type, please manually specify
the problem_type parameter during predictor init (You may specify problem_type
as one of: ['binary', 'multiclass', 'regression'])
Using Feature Generators to preprocess the data …
Fitting AutoMLPipelineFeatureGenerator…
        Available Memory:                    130450.71 MB
        Train Data (Original)  Memory Usage: 10.04 MB (0.0% of available memory)
        Inferring data type of each feature based on column values. Set
feature_metadata_in to manually specify special dtypes of the features.
        Stage 1 Generators:
                Fitting AsTypeFeatureGenerator…
                        Note: Converting 2 features to boolean dtype as they
only contain 2 unique values.
        Stage 2 Generators:
                Fitting FillNaFeatureGenerator…
        Stage 3 Generators:
                Fitting IdentityFeatureGenerator…
```

```
        Stage 4 Generators:
                Fitting DropUniqueFeatureGenerator…
        Stage 5 Generators:
                Fitting DropDuplicatesFeatureGenerator…
Training model for location C…
Train data sample weight sum: 24492.000000000004
Train data number of rows: 24492
Test data sample weight sum: 1579
Test data number of rows: 1579
        Useless Original Features (Count: 2): ['elevation:m', 'location']
                These features carry no predictive signal and should be manually
investigated.
                This is typically a feature which has the same value for all
rows.
                These features do not need to be present at inference time.
        Types of features in original data (raw dtype, special dtypes):
                ('float', []) : 42 | ['absolute_humidity_2m:gm3',
'air_density_2m:kgm3', 'ceiling_height_agl:m', 'clear_sky_energy_1h:J',
'clear_sky_rad:W', …]
                ('int', [])   :  1 | ['estimated_diff_hours']
        Types of features in processed data (raw dtype, special dtypes):
                ('float', [])     : 40 | ['absolute_humidity_2m:gm3',
'air_density_2m:kgm3', 'ceiling_height_agl:m', 'clear_sky_energy_1h:J',
'clear_sky_rad:W', …]
                ('int', [])       :  1 | ['estimated_diff_hours']
                ('int', ['bool']) :  2 | ['is_day:idx', 'is_in_shadow:idx']
        0.1s = Fit runtime
        43 features in original data used to generate 43 features in processed
data.
        Train Data (Processed) Memory Usage: 8.08 MB (0.0% of available memory)
Data preprocessing and feature engineering runtime = 0.18s …
AutoGluon will gauge predictive performance using evaluation metric:
'mean_absolute_error'
        This metric's sign has been flipped to adhere to being higher_is_better.
The metric score can be multiplied by -1 to get the metric value.
        To change this, specify the eval_metric parameter of Predictor()
Automatically generating train/validation split with holdout_frac=0.1, Train
Rows: 22042, Val Rows: 2450
User-specified model hyperparameters to be fit:
{
        'NN_TORCH': {},
        'GBM': [{'extra_trees': True, 'ag_args': {'name_suffix': 'XT'}}, {},
'GBMLarge'],
        'CAT': {},
        'XGB': {},
        'FASTAI': {},
        'RF': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
```

```
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'XT': [{'criterion': 'gini', 'ag_args': {'name_suffix': 'Gini',
'problem_types': ['binary', 'multiclass']}}, {'criterion': 'entropy', 'ag_args':
{'name_suffix': 'Entr', 'problem_types': ['binary', 'multiclass']}},
{'criterion': 'squared_error', 'ag_args': {'name_suffix': 'MSE',
'problem_types': ['regression', 'quantile']}}],
        'KNN': [{'weights': 'uniform', 'ag_args': {'name_suffix': 'Unif'}},
{'weights': 'distance', 'ag_args': {'name_suffix': 'Dist'}}],
}
Fitting 11 L1 models …
Fitting model: KNeighborsUnif … Training model for up to 1799.82s of the
1799.82s of remaining time.
        -30.0314         = Validation score   (-mean_absolute_error)
        0.04s    = Training   runtime
        0.03s    = Validation runtime
Fitting model: KNeighborsDist … Training model for up to 1799.74s of the
1799.74s of remaining time.
        -30.0815         = Validation score   (-mean_absolute_error)
        0.03s    = Training   runtime
        0.03s    = Validation runtime
Fitting model: LightGBMXT … Training model for up to 1799.66s of the 1799.66s
of remaining time.

[1000]  valid_set's l1: 17.286
[2000]  valid_set's l1: 16.8066
[3000]  valid_set's l1: 16.5904
[4000]  valid_set's l1: 16.4733
[5000]  valid_set's l1: 16.4109
[6000]  valid_set's l1: 16.3809
[7000]  valid_set's l1: 16.3455
[8000]  valid_set's l1: 16.3286
[9000]  valid_set's l1: 16.3022
[10000] valid_set's l1: 16.2884

        -16.2876         = Validation score   (-mean_absolute_error)
        14.45s   = Training   runtime
        0.19s    = Validation runtime
Fitting model: LightGBM … Training model for up to 1784.76s of the 1784.75s of
remaining time.

[1000]  valid_set's l1: 17.3864

        -17.2471         = Validation score   (-mean_absolute_error)
        3.11s    = Training   runtime
        0.03s    = Validation runtime
Fitting model: RandomForestMSE … Training model for up to 1781.57s of the
1781.56s of remaining time.
```

```
        -18.3698         = Validation score   (-mean_absolute_error)
        5.22s    = Training   runtime
        0.1s     = Validation runtime
Fitting model: CatBoost … Training model for up to 1776.08s of the 1776.08s of
remaining time.
        -17.1912         = Validation score   (-mean_absolute_error)
        125.01s  = Training   runtime
        0.01s    = Validation runtime
Fitting model: ExtraTreesMSE … Training model for up to 1651.02s of the
1651.02s of remaining time.
        -18.278  = Validation score   (-mean_absolute_error)
        1.11s    = Training   runtime
        0.08s    = Validation runtime
Fitting model: NeuralNetFastAI … Training model for up to 1649.65s of the
1649.65s of remaining time.
        -18.5803         = Validation score   (-mean_absolute_error)
        20.55s   = Training   runtime
        0.03s    = Validation runtime
Fitting model: XGBoost … Training model for up to 1629.04s of the 1629.03s of
remaining time.
        -17.1699         = Validation score   (-mean_absolute_error)
        24.02s   = Training   runtime
        0.22s    = Validation runtime
Fitting model: NeuralNetTorch … Training model for up to 1604.65s of the
1604.64s of remaining time.
        -17.2237         = Validation score   (-mean_absolute_error)
        79.07s   = Training   runtime
        0.04s    = Validation runtime
Fitting model: LightGBMLarge … Training model for up to 1525.53s of the
1525.53s of remaining time.

[1000]  valid_set's l1: 16.652
[2000]  valid_set's l1: 16.5555
[3000]  valid_set's l1: 16.5324
[4000]  valid_set's l1: 16.5278
[5000]  valid_set's l1: 16.5262
[6000]  valid_set's l1: 16.5256
[7000]  valid_set's l1: 16.5254
[8000]  valid_set's l1: 16.5254
[9000]  valid_set's l1: 16.5254
[10000] valid_set's l1: 16.5254

        -16.5254         = Validation score   (-mean_absolute_error)
        46.16s   = Training   runtime
        0.35s    = Validation runtime
Fitting model: WeightedEnsemble_L2 … Training model for up to 360.0s of the
1477.7s of remaining time.
        -15.5368         = Validation score   (-mean_absolute_error)
        0.45s    = Training   runtime
```

```
        0.0s     = Validation runtime
AutoGluon training complete, total runtime = 322.79s … Best model:
"WeightedEnsemble_L2"
TabularPredictor saved. To load, use: predictor =
TabularPredictor.load("AutogluonModels/submission_83_C/")
WARNING: eval_metric='pearsonr' does not support sample weights so they will be
ignored in reported metric.
Evaluation: mean_absolute_error on test data: -30.79885044646661
        Note: Scores are always higher_is_better. This metric score can be
multiplied by -1 to get the metric value.
Evaluations on test data:
{
    "mean_absolute_error": -30.79885044646661,
    "root_mean_squared_error": -63.830887679445,
    "mean_squared_error": -4074.382221945923,
    "r2": 0.7863794094990258,
    "pearsonr": 0.8938476674072768,
    "median_absolute_error": -2.830535888671875
}

Evaluation on test data:
-30.79885044646661
```

3 Submit

[17]:
```python
import pandas as pd
import matplotlib.pyplot as plt

train_data_with_dates = TabularDataset('X_train_raw.csv')
train_data_with_dates["ds"] = pd.to_datetime(train_data_with_dates["ds"])

test_data = TabularDataset('X_test_raw.csv')
test_data["ds"] = pd.to_datetime(test_data["ds"])
#test_data
```

```
Loaded data from: X_train_raw.csv | Columns = 48 / 48 | Rows = 92951 -> 92951
Loaded data from: X_test_raw.csv | Columns = 47 / 47 | Rows = 2160 -> 2160
```

[18]:
```python
test_ids = TabularDataset('test.csv')
test_ids["time"] = pd.to_datetime(test_ids["time"])
# merge test_data with test_ids
test_data_merged = pd.merge(test_data, test_ids, how="inner", right_on=["time",
 ↪"location"], left_on=["ds", "location"])

#test_data_merged
```

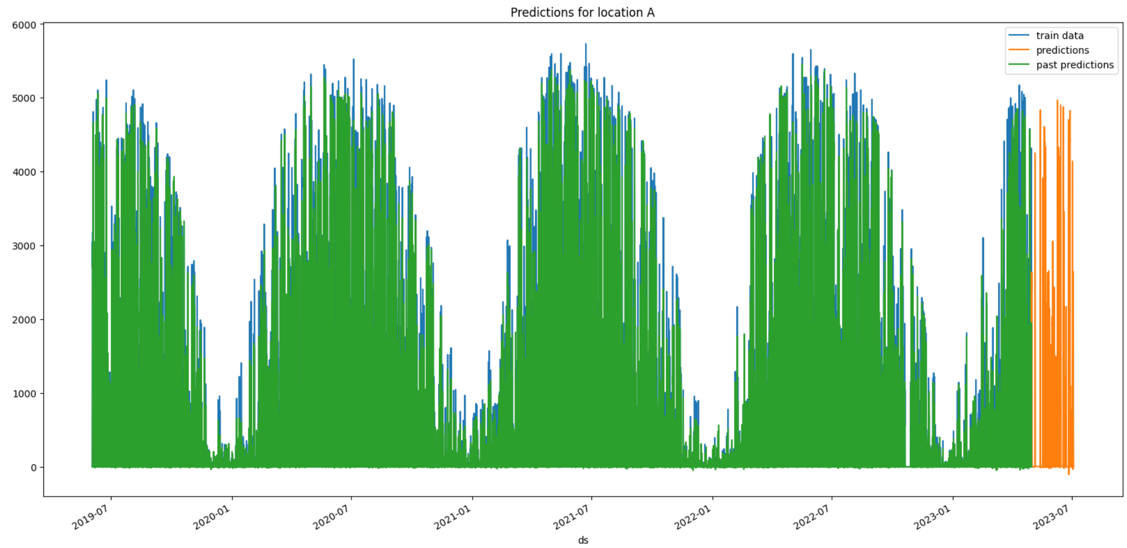
```
Loaded data from: test.csv | Columns = 4 / 4 | Rows = 2160 -> 2160
```

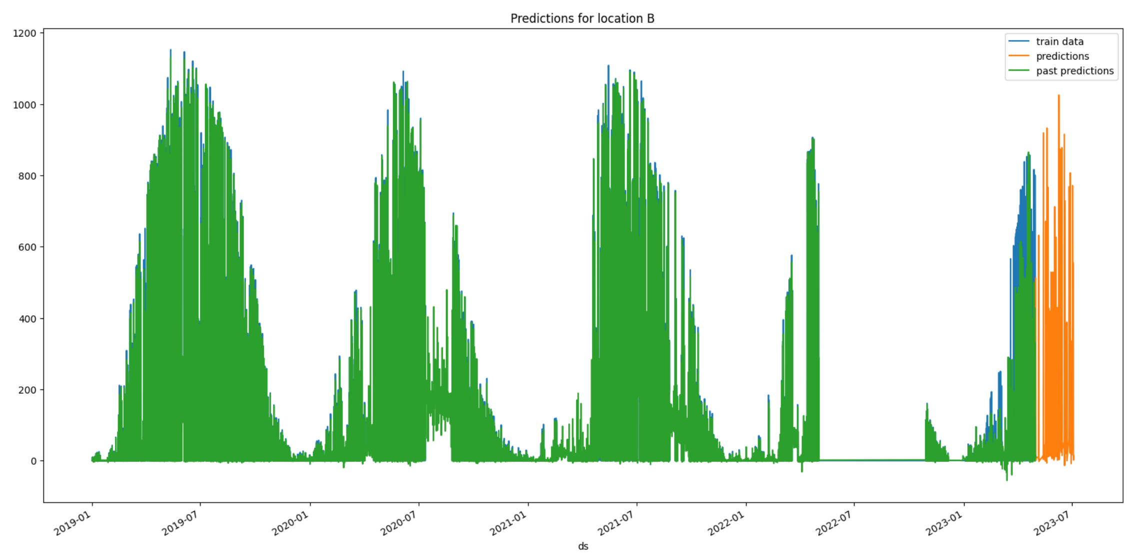
[19]:

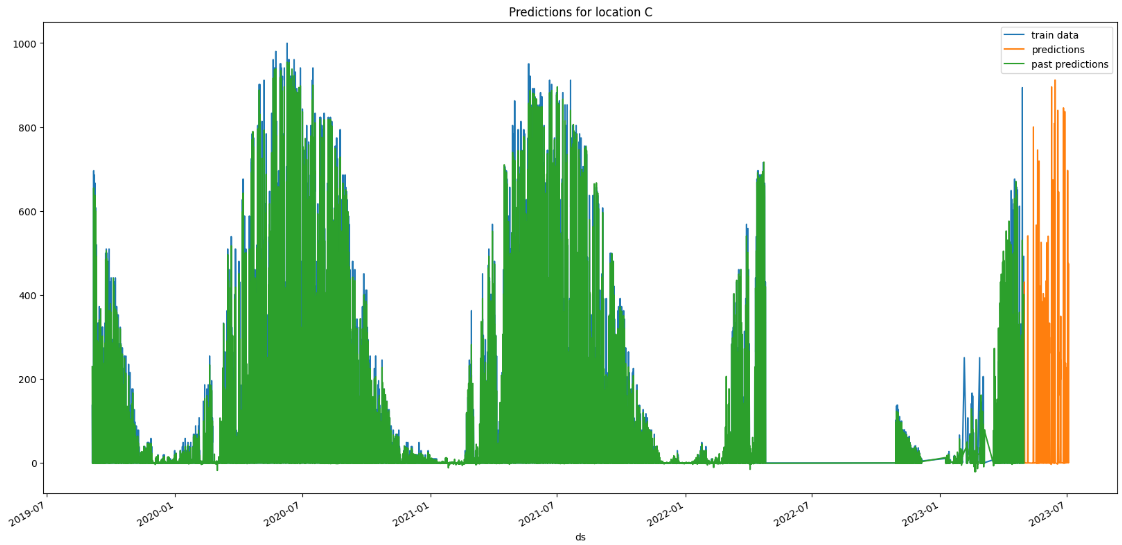
```
# predict, grouped by location
predictions = []
location_map = {
    "A": 0,
    "B": 1,
    "C": 2
}
for loc, group in test_data.groupby('location'):
    i = location_map[loc]
    subset = test_data_merged[test_data_merged["location"] == loc].
↪reset_index(drop=True)
    #print(subset)
    pred = predictors[i].predict(subset)
    subset["prediction"] = pred
    predictions.append(subset)

    # get past predictions
    past_pred = predictors[i].
↪predict(train_data_with_dates[train_data_with_dates["location"] == loc])
    train_data_with_dates.loc[train_data_with_dates["location"] == loc,␣
↪"prediction"] = past_pred
```

[20]:

```
# plot predictions for location A, in addition to train data for A
for loc, idx in location_map.items():
    fig, ax = plt.subplots(figsize=(20, 10))
    # plot train data
    train_data_with_dates[train_data_with_dates["location"]==loc].plot(x='ds',␣
↪y='y', ax=ax, label="train data")

    # plot predictions
    predictions[idx].plot(x='ds', y='prediction', ax=ax, label="predictions")

    # plot past predictions
    train_data_with_dates[train_data_with_dates["location"]==loc].plot(x='ds',␣
↪y='prediction', ax=ax, label="past predictions")

    # title
    ax.set_title(f"Predictions for location {loc}")
```


Predictions for location A
train data
predictions
past predictions
6000
5000
4000
3000
2000
1000
0
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
ds


Predictions for location B
train data
predictions
past predictions
1200
1000
800
600
400
200
0
2019-01
2019-07
2020-01
2020-07
2021-01
2021-07
2022-01
2022-07
2023-01
2023-07
ds



[21]:
```python
# concatenate predictions
submissions_df = pd.concat(predictions)
submissions_df = submissions_df[["id", "prediction"]]
submissions_df
```

[21]:
```
       id  prediction
0       0   -2.093352
1       1   -1.161028
2       2    1.035955
3       3   39.386139
4       4  346.093262
..    ...         ...
715  2155   50.987034
716  2156   33.331429
717  2157   10.217902
718  2158    2.306329
719  2159    1.580828

[2160 rows x 2 columns]
```

[22]:
```python
# Save the submission DataFrame to submissions folder, create new name based on last submission, format is submission_<last_submission_number + 1>.csv

# Save the submission
print(f"Saving submission to submissions/{new_filename}.csv")
submissions_df.to_csv(os.path.join('submissions', f"{new_filename}.csv"), index=False)
print("jall1a")
```

```
Saving submission to submissions/submission_83.csv
jall1a
```

[23]:
```
# save this running notebook
from IPython.display import display, Javascript
import time

# hei123

display(Javascript("IPython.notebook.save_checkpoint();"))

time.sleep(3)
```

```
<IPython.core.display.Javascript object>
```

[24]:
```
# save this notebook to submissions folder
import subprocess
import os
subprocess.run(["jupyter", "nbconvert", "--to", "pdf", "--output", os.path.
 ↪join('notebook_pdfs', f"hei.pdf"), "autogluon_each_location.ipynb"])
```

```
[NbConvertApp] Converting notebook autogluon_each_location.ipynb to pdf
/opt/conda/lib/python3.10/site-packages/nbconvert/utils/pandoc.py:51:
RuntimeWarning: You are using an unsupported version of pandoc (2.9.2.1).
Your version must be at least (2.14.2) but less than (4.0.0).
Refer to https://pandoc.org/installing.html.
Continuing with doubts…
  check_pandoc_version()
[NbConvertApp] Writing 110252 bytes to notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', 'notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 89019 bytes to notebook_pdfs/hei.pdf
```

[24]:
```
CompletedProcess(args=['jupyter', 'nbconvert', '--to', 'pdf', '--output',
'notebook_pdfs/hei.pdf', 'autogluon_each_location.ipynb'], returncode=0)
```

[25]:
```
# feature importance
location="A"
split_time = pd.Timestamp("2022-10-28 22:00:00")
estimated = train_data_with_dates[train_data_with_dates["ds"] >= split_time]
estimated = estimated[estimated["location"] == location]
predictors[0].feature_importance(feature_stage="original", data=estimated,␣
 ↪time_limit=60*10)
```

```
These features in provided data are not utilized by the predictor and will be
```

```
ignored: ['ds', 'elevation:m', 'sample_weight', 'location', 'prediction']
Computing feature importance via permutation shuffling for 43 features using
4394 rows with 10 shuffle sets… Time limit: 600s…
        602.15s = Expected runtime (60.21s per shuffle set)
        311.81s = Actual runtime (Completed 10 of 10 shuffle sets)
```

[25]:

	importance	stddev	p_value	n \
direct_rad:W	1.517037e+02	2.289968	3.273217e-18	10
clear_sky_rad:W	8.582451e+01	1.703378	3.841559e-17	10
diffuse_rad:W	7.504888e+01	2.161663	1.095237e-15	10
sun_azimuth:d	5.999013e+01	3.158821	2.479818e-13	10
sun_elevation:d	4.082302e+01	1.096936	5.863730e-16	10
direct_rad_1h:J	2.890493e+01	0.671130	1.575765e-16	10
clear_sky_energy_1h:J	2.494764e+01	1.374852	3.732751e-13	10
diffuse_rad_1h:J	1.564581e+01	0.809628	2.121050e-13	10
effective_cloud_cover:p	1.410651e+01	0.829709	6.696896e-13	10
total_cloud_cover:p	1.408536e+01	0.635340	6.176382e-14	10
wind_speed_u_10m:ms	1.007212e+01	1.198845	3.665292e-10	10
cloud_base_agl:m	7.394463e+00	0.472747	1.415201e-12	10
is_day:idx	6.217252e+00	0.317781	1.898621e-13	10
snow_water:kgm2	6.144440e+00	0.759398	5.122269e-10	10
fresh_snow_24h:cm	5.312308e+00	0.572670	1.516854e-10	10
relative_humidity_1000hPa:p	5.234821e+00	0.633852	4.268981e-10	10
visibility:m	5.207509e+00	0.492373	4.705255e-11	10
pressure_50m:hPa	4.911803e+00	0.681923	1.437898e-09	10
is_in_shadow:idx	4.868039e+00	0.260064	2.823661e-13	10
ceiling_height_agl:m	4.435441e+00	0.531354	3.882061e-10	10
pressure_100m:hPa	4.334008e+00	0.611214	1.652148e-09	10
wind_speed_10m:ms	4.199136e+00	0.568926	1.158211e-09	10
wind_speed_v_10m:ms	4.010390e+00	0.698641	1.065526e-08	10
sfc_pressure:hPa	3.930014e+00	0.564905	1.955842e-09	10
msl_pressure:hPa	3.814023e+00	0.395453	1.071865e-10	10
air_density_2m:kgm3	1.902994e+00	0.710650	7.007531e-06	10
t_1000hPa:K	1.902247e+00	1.067549	1.598663e-04	10
estimated_diff_hours	1.826127e+00	0.185571	8.958305e-11	10
fresh_snow_6h:cm	1.756706e+00	0.210241	3.848165e-10	10
fresh_snow_12h:cm	1.626520e+00	0.309084	2.282256e-08	10
super_cooled_liquid_water:kgm2	1.439939e+00	0.356188	2.241701e-07	10
snow_depth:cm	1.377273e+00	0.412376	1.133366e-06	10
precip_5min:mm	1.014655e+00	0.413575	1.413010e-05	10
dew_point_2m:K	9.633977e-01	0.399315	1.613398e-05	10
fresh_snow_3h:cm	9.372222e-01	0.194065	4.825074e-08	10
dew_or_rime:idx	6.006628e-01	0.177677	1.023541e-06	10
precip_type_5min:idx	5.757253e-01	0.247069	2.120778e-05	10
fresh_snow_1h:cm	4.291272e-01	0.229121	1.113596e-04	10
rain_water:kgm2	1.917926e-01	0.115546	2.641963e-04	10
prob_rime:p	1.070263e-01	0.155762	2.892224e-02	10

absolute_humidity_2m:gm3	2.790905e-02	0.160047	2.973783e-01	10
wind_speed_w_1000hPa:ms	-1.085141e-10	0.000000	5.000000e-01	10
snow_melt_10min:mm	-1.138008e-01	0.137336	9.861026e-01	10

	p99_high	p99_low
direct_rad:W	1.540571e+02	1.493504e+02
clear_sky_rad:W	8.757506e+01	8.407397e+01
diffuse_rad:W	7.727039e+01	7.282736e+01
sun_azimuth:d	6.323642e+01	5.674385e+01
sun_elevation:d	4.195033e+01	3.969571e+01
direct_rad_1h:J	2.959464e+01	2.821521e+01
clear_sky_energy_1h:J	2.636056e+01	2.353472e+01
diffuse_rad_1h:J	1.647785e+01	1.481376e+01
effective_cloud_cover:p	1.495919e+01	1.325383e+01
total_cloud_cover:p	1.473829e+01	1.343243e+01
wind_speed_u_10m:ms	1.130416e+01	8.840079e+00
cloud_base_agl:m	7.880300e+00	6.908627e+00
is_day:idx	6.543833e+00	5.890672e+00
snow_water:kgm2	6.924864e+00	5.364017e+00
fresh_snow_24h:cm	5.900834e+00	4.723782e+00
relative_humidity_1000hPa:p	5.886223e+00	4.583418e+00
visibility:m	5.713515e+00	4.701503e+00
pressure_50m:hPa	5.612607e+00	4.210999e+00
is_in_shadow:idx	5.135303e+00	4.600774e+00
ceiling_height_agl:m	4.981508e+00	3.889375e+00
pressure_100m:hPa	4.962146e+00	3.705871e+00
wind_speed_10m:ms	4.783815e+00	3.614458e+00
wind_speed_v_10m:ms	4.728375e+00	3.292405e+00
sfc_pressure:hPa	4.510561e+00	3.349468e+00
msl_pressure:hPa	4.220426e+00	3.407621e+00
air_density_2m:kgm3	2.633321e+00	1.172667e+00
t_1000hPa:K	2.999354e+00	8.051388e-01
estimated_diff_hours	2.016836e+00	1.635417e+00
fresh_snow_6h:cm	1.972768e+00	1.540643e+00
fresh_snow_12h:cm	1.944162e+00	1.308877e+00
super_cooled_liquid_water:kgm2	1.805989e+00	1.073889e+00
snow_depth:cm	1.801067e+00	9.534785e-01
precip_5min:mm	1.439682e+00	5.896286e-01
dew_point_2m:K	1.373769e+00	5.530260e-01
fresh_snow_3h:cm	1.136661e+00	7.377835e-01
dew_or_rime:idx	7.832597e-01	4.180659e-01
precip_type_5min:idx	8.296352e-01	3.218155e-01
fresh_snow_1h:cm	6.645924e-01	1.936620e-01
rain_water:kgm2	3.105377e-01	7.304745e-02
prob_rime:p	2.671008e-01	-5.304827e-02
absolute_humidity_2m:gm3	1.923874e-01	-1.365693e-01
wind_speed_w_1000hPa:ms	-1.085141e-10	-1.085141e-10

```
snow_melt_10min:mm                2.733802e-02 -2.549395e-01
```

[ ]:
```
# feature importance
observed = train_data_with_dates[train_data_with_dates["ds"] < split_time]
observed = observed[observed["location"] == location]
predictors[0].feature_importance(feature_stage="original", data=observed,␣
↪time_limit=60*10)
```

```
These features in provided data are not utilized by the predictor and will be
ignored: ['ds', 'elevation:m', 'sample_weight', 'location', 'prediction']
Computing feature importance via permutation shuffling for 43 features using
5000 rows with 10 shuffle sets… Time limit: 600s…
        628.12s = Expected runtime (62.81s per shuffle set)
```

[ ]:
```
display(Javascript("IPython.notebook.save_checkpoint();"))
time.sleep(3)

subprocess.run(["jupyter", "nbconvert", "--to", "pdf", "--output", os.path.
↪join('notebook_pdfs', f"{new_filename}_with_feature_importance.pdf"),␣
↪"autogluon_each_location.ipynb"])
```

[ ]:
```
# import subprocess

# def execute_git_command(directory, command):
#     """Execute a Git command in the specified directory."""
#     try:
#         result = subprocess.check_output(['git', '-C', directory] + command,␣
↪stderr=subprocess.STDOUT)
#         return result.decode('utf-8').strip(), True
#     except subprocess.CalledProcessError as e:
#         print(f"Git command failed with message: {e.output.decode('utf-8').
↪strip()}")
#         return e.output.decode('utf-8').strip(), False

# git_repo_path = "."

# execute_git_command(git_repo_path, ['config', 'user.email',␣
↪'henrikskog01@gmail.com'])
# execute_git_command(git_repo_path, ['config', 'user.name', hello if hello is␣
↪not None else 'Henrik eller Jørgen'])

# branch_name = new_filename

# # add datetime to branch name
# branch_name += f"_{pd.Timestamp.now().strftime('%Y-%m-%d_%H-%M-%S')}"

# commit_msg = "run result"
```

```
# execute_git_command(git_repo_path, ['checkout', '-b',branch_name])

# # Navigate to your repo and commit changes
# execute_git_command(git_repo_path, ['add', '.'])
# execute_git_command(git_repo_path, ['commit', '-m',commit_msg])

# # Push to remote
# output, success = execute_git_command(git_repo_path, ['push', 'origin',branch_name])

# # If the push fails, try setting an upstream branch and push again
# if not success and 'upstream' in output:
#     print("Attempting to set upstream and push again...")
#     execute_git_command(git_repo_path, ['push', '--set-upstream', 'origin',branch_name])
#     execute_git_command(git_repo_path, ['push', 'origin', 'henrik_branch'])

# execute_git_command(git_repo_path, ['checkout', 'main'])
```